Journal of Multivariate Analysis 102 (2011) 1166–1174

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Semi-varying coefficient models with a diverging number of components

Gaorong Li [a], Liugen Xue [a], Heng Lian [b,*]

[a] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China

[b] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

ARTICLE INFO

Article history:
Received 17 November 2010
Available online 7 April 2011

AMS subject classifications:
62G08
62G20

Keywords:
B-spline basis
Diverging parameters
Semi-varying coefficient models

ABSTRACT

Semiparametric models with both nonparametric and parametric components have become increasingly useful in many scientific fields, due to their appropriate representation of the trade-off between flexibility and efficiency of statistical models. In this paper we focus on semi-varying coefficient models (a.k.a. varying coefficient partially linear models) in a "large n, diverging p" situation, when both the number of parametric and nonparametric components diverges at appropriate rates, and we only consider the case $p = o(n)$. Consistency of the estimator based on B-splines and asymptotic normality of the linear components are established under suitable assumptions. Interestingly (although not surprisingly) our analysis shows that the number of parametric components can diverge at a faster rate than the number of nonparametric components and the divergence rates of the number of the nonparametric components constrain the allowable divergence rates of the parametric components, which is a new phenomenon not established in the existing literature as far as we know. Finally, the finite sample behavior of the estimator is evaluated by some Monte Carlo studies.

1. Introduction

Regression problems, where one is interested in characterizing the relationships between a set of covariates $X_1, \ldots, X_p$ and a response variable Y, are fundamental in statistics. With modern technologies producing data sets with large size and dimensions, there has been considerable interest in investigating the "diverging p" asymptotic framework. This can be traced back at least to the study of consistency and asymptotic normality of M-estimators with a diverging number of predictors [12,25,16,17,23]. In many real applications, however, parametric models are not flexible enough to capture the true underlying relationships between covariates and response. Relaxation of the parametric assumptions leads to various proposals for semiparametric models, including partially linear models [5], additive models [6], varying coefficient models [7] and some hybrids.

More recently, investigation of models with a large number of parameters is revived by the incorporation of variable selection using penalization [20,3,28,26,29]. The works on variable selection with a diverging number of parameters mostly focus on parametric models [4,9,27,14]. [24] extended the framework to partially linear models. For additive models, several independent works [18,15,10] considered the diverging p case. In contrast to partially linear models where the number of parameters in the linear part diverges, for additive models the number of nonparametric components diverges to infinity with sample size.

* Corresponding author.
E-mail address: henglian@ntu.edu.sg (H. Lian).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.03.010

Here we study the semi-varying coefficient models with the number of parametric and nonparametric components both diverging to infinity at suitable rates. This situation has not been considered in the literature to the best of our knowledge. More specifically, suppose we have n i.i.d. observations from the model

$$Y_i = \sum_{j=1}^{p_1} W_{ij}\alpha_j(t_i) + \sum_{j=1}^{p_2} X_{ij}\beta_j + \epsilon_i = W_i^T\alpha(t_i) + X_i^T\beta + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where t is the index variable, W_{ij}, X_{ij} are the covariates, $\alpha = (\alpha_1, \ldots, \alpha_{p_1})^T$ are the varying coefficients, $\beta = (\beta_1, \ldots, \beta_{p_2})^T$ are the linear coefficients, and the number of nonparametric and parametric components p_1, p_2 implicitly depends on sample size n. We will denote the true parameters by α_0 and β_0. In [13], the authors considered generalized semi-varying coefficient models for the fix p case. Some interesting questions arise in the case of diverging number of parameters. It is naturally expected that the number of parametric components can diverge faster than the number of nonparametric components, but exactly how? And does the present of nonparametric components pose some constraint on the possible divergence rate of the number of parametric components, or vice versa? Here we provide a partial answer to these questions based on consistency and asymptotic normality analysis, using B-spline estimators for the nonparametric components.

When the covariates satisfy certain conditions (see Condition (c1) in Section 2), we show that if $p_1/n^{2d/(2d+1)} \to 0$ and $p_2/n \to 0$, then the estimation is consistent, where d is the smoothness parameter of the nonparametric varying coefficients. Note that this implies the number of parametric components can diverge at a faster rate. Besides, if $p_1p_2/n^{(d-1/2)/(2d+1)} \to 0$, we can show the asymptotic normality of the parametric part. This condition characterizes the *multiplicative* effect of the presence of the nonparametric components on the allowable number of parametric components. Precise additional assumptions required for these results will be explained next.

2. Estimation and asymptotics

In this paper, we use B-splines to approximate the varying coefficients $\alpha_j(t)$, $1 \le j \le p_1$. For simplicity, we assume the index variable t in (1) has a distribution supported on the interval $[0, 1]$. To approximate a function on $[0, 1]$, we partition the interval $[0, 1]$ into K' subintervals $[(k-1)/K', k/K']$, for $k = 1, 2, \ldots, K'$ with $K' = K'(n)$ being a sequence of natural numbers diverging to infinity as sample size n goes to infinity. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q-1$ and globally $q-2$ times differentiable. The collection of such polynomial splines has a normalized B-spline basis $\{B_1(t), \ldots, B_K(t)\}$ with $K = K' + q$. As in [2], the basis satisfies (i) $B_k \ge 0$, $k = 1, \ldots, K$, (ii) $\sum_{k=1}^{K} B_k(t) \equiv 1$ and (iii) B_j is supported inside an interval of length q/K and at most q of the basis functions are nonzero at any given t. Using spline expansions, we can approximate the coefficients by $\alpha_j(t) \approx \sum_k a_{jk}B_k(t)$. It is also possible to construct irregular subintervals based on observed values of the index variable, or to specify different K for different varying coefficient, but we make the above choices for simplicity.

Using B-splines expansion, we can estimate the model (1) using the simple least squares procedure which can be solved in closed form,

$$(\hat{a}, \hat{\beta}) = \arg\min_{(a,\beta)} \frac{1}{2} \sum_i \left(Y_i - \sum_{j=1}^{p_1}\sum_{k=1}^{K} W_{ij}a_{jk}B_k(t_i) - \sum_{j=1}^{p_2} X_{ij}\beta_j \right)^2. \tag{2}$$

Using the notation

$$Z_j = \begin{pmatrix} W_{1j}B_1(t_1) & W_{1j}B_2(t_1) & \cdots & W_{1j}B_K(t_1) \\ \cdots & \cdots & \cdots & \cdots \\ W_{nj}B_1(t_n) & W_{nj}B_2(t_n) & \cdots & W_{nj}B_K(t_n) \end{pmatrix}_{n\times K},$$

$Z = (Z_1, \ldots, Z_{p_1})$, $X = (X_1, \ldots, X_{p_2})$, $Y = (Y_1, \ldots, Y_n)^T$, (2) can be written in matrix form as

$$\arg\min_{(a,\beta)} \frac{1}{2}\|Y - Za - X\beta\|^2. \tag{3}$$

In practice, we need to choose some parameters including the spline order q and the number of basis K. As a commonly adopted strategy, we fix $q = 4$ (cubic splines) and choose K using generalized cross-validation (GCV).

Next we will introduce additional definitions and notations for our asymptotic analysis. Let $w = (w_1, \ldots, w_{p_1})^T$, $x = (x_1, \ldots, x_{p_2})^T$ be the random variables representing the predictors. Also let $\mathcal{G}$ denote the subspace of functions on $R^{p_1} \times [0, 1]$,

$$\mathcal{G} := \left\{ g(w, t) : g(w, t) = w^T h(t), h(t) = (h_1(t), \ldots, h_{p_1}(t))^T \text{with some functions } h_j(t) \text{ and } E\sum_{j=1}^{p_1} w_j^2 h_j^2(t) < \infty \right\},$$

and for any random variable x with $E(x^2) < \infty$, let $E_{\mathcal{G}}(x)$ denote the projection of x onto $\mathcal{G}$ in the sense that

$$E\{(x - E_{\mathcal{G}}(x))(x - E_{\mathcal{G}}(x))\} = \inf_{g\in\mathcal{G}} E\{(x - g(w, t))(x - g(w, t))\}.$$

Definition of $E_{\mathcal{G}}(x)$ trivially extends to the case when x is a random vector by componentwise projection.

In the theoretical studies of our estimator, we will use the following decomposition for $x \in R^{p_2}$,

$$x = \theta(w,t) + u = \theta(w,t) - g(w,t) + g(w,t) + u, \tag{4}$$

with $\theta(w,t) = E(x|w,t), g(w,t) = E_{\mathcal{G}}(x)$. Note that since the conditional expectation $E(x|w,t)$ can be interpreted as projection onto the space $\{h(w,t), Eh^2 < \infty\}$ of which $\mathcal{G}$ is a subspace, we see that we also have $g(w,t) = E_{\mathcal{G}}(\theta(w,t))$. Let $\Xi = E\{(x - g(w,t))(x - g(w,t))^T\}$ which can be considered as the residual variance of x after removing its projection onto $\mathcal{G}$.

Now we can state the assumptions used in our asymptotic results.

(c1) The covariates have finite fourth moments, $\max_j EW_{ij}^4 < \infty$, $\max_j EX_{ij}^4 < \infty$, and the eigenvalues of $E\{(w^T, x^T)^T(w^T, x^T)\}$ are bounded away from zero and infinity.
(c2) The noises ϵ_i are independent of covariates, have mean zero, variance σ^2, and finite fourth moment.
(c3) The index variable t has a density bounded away from 0 and infinity on $[0, 1]$.
(c4) For $1 \le j \le p_1$, $\alpha_{0j}(t)$ satisfies a Lipschitz condition of order $d > 1/2$: $|\alpha_{0j}^{(\lfloor d \rfloor)}(t) - \alpha_{0j}^{(\lfloor d \rfloor)}(s)| \le C|s-t|^{d-\lfloor d \rfloor}$, where $\lfloor d \rfloor$ is the biggest integer strictly smaller than d and $\alpha_{0j}^{(\lfloor d \rfloor)}(t)$ is the $\lfloor d \rfloor$-th derivative of $\alpha_{0j}(t)$. The order of the B-spline used satisfies $q \ge d + 2$.
(c5) $Kp_1/n \to 0, p_1/K^{2d} \to 0, p_2/n \to 0$.
(c6) The eigenvalues of Ξ are bounded away from zero and infinity.
(c7) In the decomposition (4), each component of $g(w,t)$ can be written in the form $\sum_{j=1}^{p_1} w_j h_j(t)$ for some h_j. We assume all h_j satisfy a Lipschitz condition of order $d_g > 1/2$: $|h_j^{(\lfloor d_g \rfloor)}(t) - h_j^{(\lfloor d_g \rfloor)}(s)| \le C|s-t|^{d_g - \lfloor d_g \rfloor}$. The order of the B-spline used satisfies $q \ge d_g + 2$.
(c8) $np_1^2/K^{2(d+d_g)} \to 0, p_1p_2K/n \to 0, p_1^2/K^{2d-1} \to 0, p_1p_2/K^{2d_g} \to 0$.
(c9) $p_2^2/n \to 0, p_1^2p_2^2/K^{2d-1} \to 0, np_1^2p_2^2/K^{2(d+d_g)} \to 0$.

In Condition (c1), we require that the eigenvalues of the second moment matrix of covariates are bounded away from zero and infinity. Similar assumptions are used in other papers including [22]. It is possible to relax this to require that the minimum eigenvalue is bounded away from zero with some rates converging to zero, but at the cost of more complicated statements in the theorems below. Conditions (c2)–(c4) are standard. The convergence rate (5) would be void without Condition (c5). (c6)–(c8) are used in showing the faster convergence rate of the parametric component. (c6) and (c7) imply that x is not in $\mathcal{G}$ and its projection onto $\mathcal{G}$ is smooth enough. Assumption(c6) can be seen as a natural extension of the assumption of nonsigularity of $E(X^TX)$ in linear regression models, which guarantees the identifiability of the model. Obviously if the covariates for the parametric part is contained in $\mathcal{G}$, then the coefficients are not estimable. These conditions are similar to assumption (A2) and Condition 1 in [24] respectively for high-dimensional partially linear models.

Theorem 1 (*Convergence Rates*). *Under conditions* (c1)–(c5), *the minimizer of* (3), $(\hat{a}, \hat{\beta})$, *satisfies*

$$\|\hat{a} - a_0\|^2 + K\|\hat{\beta} - \beta_0\|^2 = O\left(\frac{K^2p_1}{n} + \frac{Kp_2}{n} + \frac{p_1}{K^{2d-1}}\right),$$

where a_0 *is any vector satisfying* $\|\alpha_{0j}(t) - \sum_k a_{0jk}B_k(t)\| = O(K^{-2d})$. *As an immediate corollary,*

$$\sum_{j=1}^{p_1}\|\hat{\alpha}_j(t) - \alpha_{0j}(t)\|^2 + \sum_{j=1}^{p_2}(\hat{\beta}_j - \beta_{0j})^2 = O\left(\frac{Kp_1}{n} + \frac{p_2}{n} + \frac{p_1}{K^{2d}}\right), \tag{5}$$

where $\alpha_{0j}(t)$ *denotes the true coefficients and* $\hat{\alpha}_j(t) = \sum_k \hat{a}_{jk}B_k(t)$.

For the parametric part, under additional assumptions (c6)–(c8), we have the faster rate

$$\sum_{j=1}^{p_2}(\hat{\beta}_j - \beta_{0j})^2 = O\left(\frac{p_2}{n}\right).$$

Remark 1. Based on the first and the third term in the convergence rate (5), we see that $K \sim n^{1/(2d+1)}$ is the optimal choice, which is the same as in the fixed p case [19]. Although in the formulation (2), the regression coefficients a_{jk} and β_j appear to be in the equal footing, they behave radically differently, due to the face that we are concerned with estimation of α_j, which consists of approximation error $\|\alpha_j - \sum_k a_{0jk}B_k\|^2$ as well as estimation error $\sum_k(a_{jk} - a_{0jk})^2$. It is necessary that $K \to \infty$ to reduce approximation error, at the cost of larger estimation error.

Remark 2. The convergence rate (5) concerns the sum of the nonparametric part and the parametric part. This is of course related to the prediction error. The rate $Kp_1/n + p_2/n + p_1/K^{2d}$ is seen to be a sum of the rates for the nonparametric part ($Kp_1/n + p_1/K^{2d}$) and the parametric part (p_2/n). Indeed, the second part of the theorem shows that the rate for the parametric part is p_2/n. As seen in the Appendix, the proof strategy is to profile out the nonparametric part in determining

the rates for the parametric part. We expect that by following the same strategy, we can show the convergence rate for the nonparametric part is indeed $Kp_1/n + p_1/K^{2d}$. This seems however quite messy and we skip this derivation. Note that in the classical case where p_1 and p_2 are bounded (or slightly more generally when p_1 and p_2 are of the same order), the term p_2/n is of smaller order than Kp_1/n, and thus (5) directly implies $\sum_j \|\hat{\alpha}_j - \alpha_{0j}\|^2 = O(Kp_1/n + p_1/K^{2d})$.

Under the stronger assumption (c9), we can show asymptotic normality of the parametric components.

Theorem 2 (*Asymptotic Normality of Parametric Part*)**.** *Let A_n be a deterministic $m \times p_2$ matrix with m an integer that does not change with n, and $\Sigma_n = A_n \Xi^{-1} A_n^T$ (Ξ is defined below (4)). Under conditions* (c1)–(c9)*,*

$$\sqrt{n}\Sigma_n^{-1/2} A_n(\hat{\beta} - \beta_0) \to N(0, \sigma^2 I_m) \text{ in distribution},$$

where I_m is the $m \times m$ identity matrix.

It is also possible to obtain the (pointwise) asymptotic normality of the nonparametric component as below. However, as pointed out in [8,11], it is in general very difficult to estimate the bias term.

Theorem 3. *Under conditions* (c1)–(c8)*, we have*

$$\{\text{var}(\hat{\alpha}_j(t))\}^{-1/2}(\hat{\alpha}_j(t) - E(\hat{\alpha}_j(t))) \to N(0, 1) \quad \text{in distribution},$$

where var(.) *and* $E(.)$ *above denote the variance and expectation conditional on covariates respectively.*

Remark 3. Suppose we use the optimal choice $K \sim n^{1/(2d+1)}$. For consistency, the divergence rate of the number of parameters can be $p_1 = o(n^{2d/(2d+1)})$ and $p_2 = o(n)$ (this is same as Condition (c5) when optimal K is chosen). For simplicity of this discussion, assume d_g is large enough so that the assumptions involving d_g in (c8) and (c9) are satisfied. Then the conditions in (c8) and (c9) reduce to $p_1 p_2 = o(n^{(d-1/2)/(2d+1)})$, $p_2 = o(n^{1/2})$. As long as $p_1 > 0$, $p_1 p_2 = o(n^{(d-1/2)/(2d+1)})$ implies $p_2 = o(n^{(d-1/2)/(2d+1)})$ and the assumption $p_2 = o(n^{1/2})$ is trivially satisfied. The condition $p_2 = o(n^{1/2})$ is only here for completeness so that the theorem on asymptotic normality is valid even for purely linear models. Note that even with a single nonparametric component $p_1 = 1$, the possible divergence rate of p_2 is reduced from $o(n^{1/2})$ to $o(n^{(d-1/2)/(2d+1)})$.

3. Simulation

In this section we use some simulations to evaluate the finite sample performance of the semi-varying coefficient models when the number of parameters increases. The data sets are generated from model (1) with sample size $n = 100$ and noises $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.5$, 1 and 2. The index variable t is sampled uniformly on [0, 1], and the predictors are $S = (W, X)$, with $S_{i1} = 1$ and other S_{ij}s marginally standard normal within subject correlations $\text{Cov}(S_{ij_1}, S_{ij_2}) = (1/2)^{|j_1 - j_2|}$. First we set $p_1 = 2$, $p_2 = 9$, the first two coefficient functions are

$$\alpha_1(t) = 3\sin(2\pi t),$$
$$\alpha_2(t) = 8t(1-t).$$

Those 9 parameters for the parametric part are specified as $\beta_1 = \beta_2 = \beta_3 = 0.5, \beta_4 = \beta_5 = \beta_6 = 0.2$ and $\beta_7 = \beta_8 = \beta_9 = 0.1$. We also generate data sets with larger p_1, p_2 by repeating the coefficients. For example, if $p_1 = 4$, we set $\alpha_3 = \alpha_1$ and $\alpha_4 = \alpha_2$ in the simulations. For all scenarios, 100 data sets are generated and fitted, with K chosen by GCV criterion. In Table 1, we report the mean squared errors for the nonparametric and the parametric part, defined by

$$\text{MSE1} = \sum_{j=1}^{p_1} \|\hat{\alpha}_j - \alpha_j\|^2, \qquad \text{MSE2} = \sum_{j=1}^{p_2} (\hat{\beta}_j - \beta_j)^2.$$

As can be seen from the table, increasing the number of parameters (for example comparing $p_1 = 2, p_2 = 9$ with $p_1 = 4, p_2 = 18$) makes the estimation more difficult. Fig. 1 shows the estimation of α_1 with models of different dimensions. Also, the table shows that increasing the number of parameters in the nonparametric part has an adverse effect on the estimation of the linear part (for example comparing $p_1 = 4, p_2 = 36$ with $p_1 = 8, p_2 = 36$), and vice versa, as expected.

4. Conclusion

In this paper we study the estimation of semi-varying coefficient models when the number of parametric components and the number of nonparametric components both diverge with sample size. In order to achieve consistency in estimation, our asymptotic results show that the number of parametric components can diverge faster than the number of nonparametric components. For asymptotic normality, the required condition shows an interesting interaction between the two.

Variable selection by penalization has attracted much attention recently with high dimensional data, typically assuming sparsity of the model. As a future work, it is thus interesting to use penalization to identify the nonzero coefficients in semi-varying coefficient models automatically. Another possible extension of the current work is to consider generalized semi-varying coefficient models. Variable selection for such models has been considered in [13] based on local linear regression for a fixed dimension.

Table 1
Estimation errors of different scenarios based on 100 replications, with $n = 100$. MSE1 is the mean squared error for the nonparametric part and MSE2 is for the parametric part.

	p_1	p_2	MSE1	MSE2
$\sigma = 0.5$	2	9	0.056	0.047
	4	18	0.203	0.143
	8	36	1.361	0.907
	4	36	0.304	0.400
$\sigma = 1$	2	9	0.179	0.191
	4	18	0.572	0.527
	8	36	2.267	2.289
	4	36	0.731	1.465
$\sigma = 2$	2	9	0.526	0.738
	4	18	1.482	1.919
	8	36	5.502	8.337
	4	36	2.029	5.799

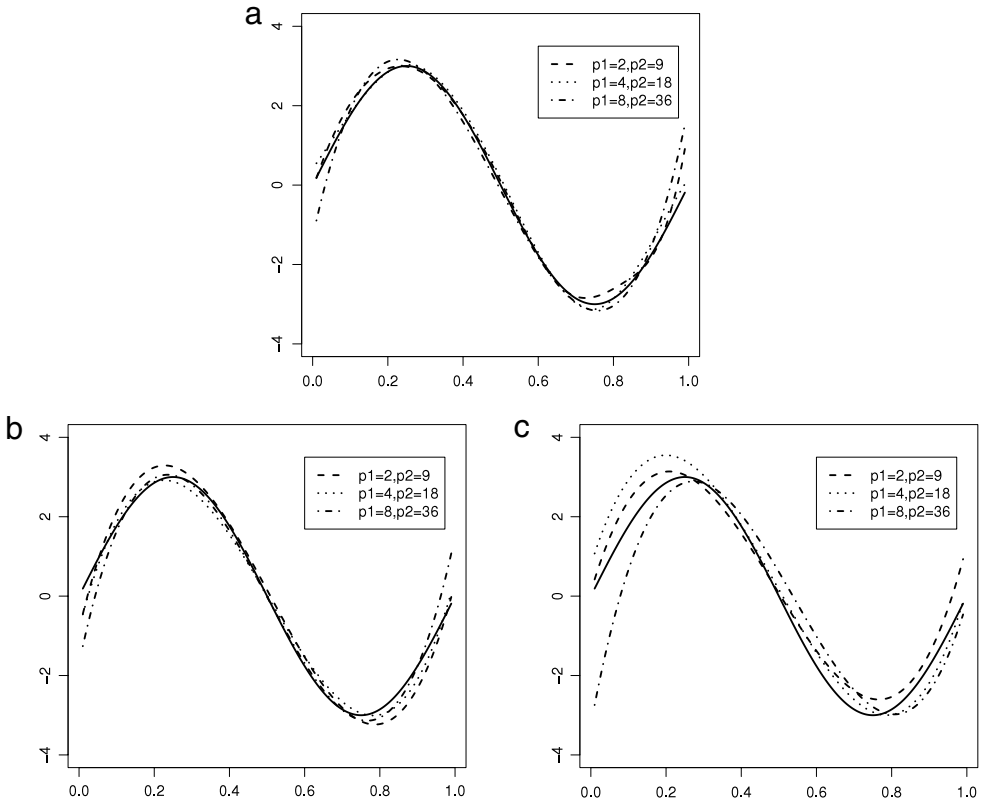

Fig. 1. Estimation of α_1 for different dimensionality, with the true α_1 shown as the solid line. (a) $\sigma = 0.5$; (b) $\sigma = 1$; (c) $\sigma = 2$.

Acknowledgments

We sincerely thank the AE and a referee for their suggestions that have improved the quality of our paper. Gaorong Li's research was supported by the Ph.D. Program Foundation of Ministry of Education of China (20101103120016), the Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the jurisdiction of Beijing Municipality (PHR20110822), the Training Programme Foundation for the Beijing Municipal Excellent Talents (2010D005015000002) and the Doctor Foundation of BJUT (X0006013201101). Liugen Xue's research was supported by the National Natural Science Foundation of China (10871013), the Natural Science Foundation of Beijing (1102008), and the Beijing Municipal Education Commission Foundation in 2011 and PHR(IHLB). The research of Heng Lian is supported by Singapore Ministry of Education Tier 1 36/09.

Appendix

Proof of Theorem 1. The convergence rate for the nonparametric component is relatively easy to show.

Suppose $\alpha_{nj}(t) = \sum_{k=1}^{K} a_{0jk}B_k(t)$ is the best approximating spline for $\alpha_{0j}(t)$ with $\|\alpha_{nj} - \alpha_{0j}\|^2 = O(K^{-2d})$. Denote $V = (Z, X/\sqrt{K})$, $\hat{b} = (\hat{a}, \sqrt{K}\hat{\beta})$ and $b_0 = (a_0, \sqrt{K}\beta_0)$. By the definition of $\hat{a}$, $\hat{\beta}$ in (2), we have

$$\begin{aligned} 0 &\geq Q(\hat{b}) - Q(b_0) \\ &= \|Y - Z\hat{a} - X\hat{\beta}\|^2/2 - \|Y - Za_0 - X\beta_0\|^2/2 \\ &= (Y - Vb_0)^T V(b_0 - \hat{b}) + \|V(b_0 - \hat{b})\|^2/2. \end{aligned}$$

Let $\eta = P_V(Y - Vb_0)$, where $P_V = V(V^TV)^{-1}V^T$, be the projection of $Y - Vb_0$ onto the columns of $V = (Z, X/\sqrt{K})$. Lemma 1 shows that $\|\eta\|^2 = O_p(Kp_1 + p_2 + np_1/K^{2d})$. Using the Cauchy–Schwartz inequality, the above displayed equation can be continued as

$$0 \geq -|O_p(Kp_1 + p_2 + np_1/K^{2d})| + \frac{1}{4}\|V(b_0 - \hat{b})\|^2. \tag{6}$$

Using now Lemma A.1 in [21] together with Condition (c1), which implies that $\|V(b_0 - \hat{b})\|^2 \sim (n/K)\|b_0 - \hat{b}\|^2 = (n/K)(\|a_0 - \hat{a}\|^2 + K\|\beta_0 - \hat{\beta}\|^2)$, (6) leads to $\|\hat{a} - a_0\|^2 + K\|\hat{\beta} - \beta_0\|^2 = O_p(K^2p_1/n + Kp_2/n + p_1/K^{2d-1})$. The convergence rate for $\sum_{j=1}^{p_1}\|\hat{\alpha}_j - \alpha_{0j}\|^2$ is obtained from the well known relation $\|\sum_k a_kB_k(t)\|^2 \sim \|a\|^2/K$ for any $a = (a_1, \ldots, a_K)$.

Now consider the faster convergence rate of the parametric components, which we show by profiling out a in (2). For any given β, let $\hat{a}(\beta)$ be the minimizer of (2) when β is fixed. It is obvious that

$$\hat{a}(\beta) = (Z^TZ)^{-1}Z^T(Y - X\beta).$$

Let β_0 be the true parameter and set $\hat{\beta} = \beta_0 + \gamma_1 u$ with $\gamma_1 = C\sqrt{p_2/n}$ for some $C > 0$, and $\|u\| = 1$. We will show that $\inf_{\|u\|=1} Q(\hat{a}(\hat{\beta}), \hat{\beta}) - Q(\hat{a}(\beta_0), \beta_0) > 0$ with probability approaching 1 for C large enough and the result will follow.

Using the closed form expression for $\hat{a}(\beta)$, we get

$$Q(\hat{a}(\hat{\beta}), \hat{\beta}) - Q(\hat{a}(\beta_0), \beta_0) = -(\tilde{Y} - \tilde{X}\beta_0)(\gamma_1\tilde{X}u) + (1/2)\|\gamma_1\tilde{X}u\|^2$$

where for any random matrix W with n rows, we set $\tilde{W} = Q_ZW = W - P_ZW$ to be the projection of columns of W onto the orthogonal complement of the column space of Z, where $P_Z = Z(Z^TZ)^{-1}Z^T$.

In Lemma 2(i), we show that $\|(\tilde{Y} - \tilde{X}\beta_0)^T(\tilde{X}u)\| = O(\sqrt{np_2})$. Since the eigenvalues of $\tilde{X}^T\tilde{X}/n$ are bounded away from zero by Lemma 2(ii) and Condition (c6), $Q(\hat{a}(\hat{\beta}), \hat{\beta}) - Q(\hat{a}(\beta_0), \beta_0)$ is bounded below by

$$nc\gamma_1^2 + O(\sqrt{np_2})\gamma_1,$$

for some $c > 0$. Thus if $\gamma_1 = C\sqrt{np_2}/n$ for $C > 0$ sufficiently large, the above displayed expression will be positive. □

Proof of Theorem 2. Since $Y = r + X\beta_0 + \epsilon$ where $r = (r_1, \ldots, r_n)$ with $r_i = \sum_{j=1}^{p_1} W_{ij}\alpha_{0j}(t)$, and denote by a_0 the vector containing the spline coefficients that achieve optimal approximation of $\alpha_j(t)$, $1 \leq j \leq p_1$, and set $v = r - Za_0$. The first order conditions for (3) are

$$\begin{aligned} -Z_j^T(\epsilon + v - Z(a - a_0) - X(\beta - \beta_0)) &= 0, \quad j = 1, \ldots, p_1, \\ -X_j^T(\epsilon + v - Z(a - a_0) - X(\beta - \beta_0)) &= 0, \quad j = 1, \ldots, p_2. \end{aligned}$$

From the first displayed equation above, we get $Z(a - a_0) = Z(Z^TZ)^{-1}Z^T(\epsilon + v - X(\beta - \beta_0))$ and plugging this into the second displayed equation above we get

$$-X_j^T(\epsilon + v - Z(Z^TZ)^{-1}Z^T(\epsilon + v - X(\beta - \beta_0)) - X(\beta - \beta_0)) = 0, \quad j = 1, \ldots, p_2,$$

that is,

$$-X_j^T(\widetilde{\epsilon + v} - \tilde{X}(\beta - \beta_0)) = 0, \quad j = 1, \ldots, p_2,$$

from which we get

$$\sqrt{n}\Sigma_n^{-1/2}A_n(\hat{\beta} - \beta_0) = \sqrt{n}\Sigma_n^{-1/2}A_n(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T(\epsilon + v).$$

By Lemma 2(ii), we can replace $(\tilde{X}^T\tilde{X}/n)^{-1}$ by Ξ^{-1} which only results in a multiplicative factor $1 + o(1)$ and thus does not disturb the asymptotic distribution.

It is easily shown

$$\left\| \frac{1}{\sqrt{n}}\Sigma_n^{-1/2}A_n\Xi^{-1} \right\| = O(\sqrt{p_2/n}).$$

Combining this with $\|\tilde{X}^T v\| = O(\sqrt{np_1^2 p_2/K^{2d-1}} + np_1\sqrt{p_2}/K^{(d+d_g)})$ (combining bounds (8)–(10) in Lemma 2(i)), and using Condition (c9), we get

$$\left\| \frac{1}{\sqrt{n}} \Sigma_n^{-1/2} A_n \Xi^{-1} \tilde{X}^T v \right\| = o(1).$$

Finally, when $p_2^2/n \to 0$, $\sqrt{n}\Sigma_n^{-1/2} A_n (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\epsilon$ can be shown to converge to $N(0, \sigma^2 I)$ by Lindeberg–Feller's central limit theorem using standard arguments. □

Proof of Theorem 3. Since we have $b = (V^TV)^{-1}V^TY$ (using notations in the proof of Theorem 1), we have $b - E(b) = (V^TV)^{-1}V^T\epsilon$. Since $\hat{\alpha}_j(t) = \sum_k \hat{a}_{jk}B_k(t)$, it is sufficient to show that for any vector c_n whose components are not all zero, we have $c_n^T(\hat{a} - E(\hat{a}))/\mathrm{var}(c_n^T(\hat{a} - E(\hat{a})))^{-1/2} \to N(0, 1)$. The demonstration of this follows exactly the same arguments of Lemma A.8 in [11]. □

Lemma 1. *Following notations defined in the proof of Theorem 1, $\|\eta\|^2 = \|P_V(Y - Vb_0)\|^2$ is of order $O(Kp_1 + p_2 + np_1/K^{2d})$.*

Proof. Denote $r_i = \sum_{j=1}^{p_1} W_{ij}\alpha_{0j}(t_i)$ and $r = (r_1, \ldots, r_n)^T$. We have $Y - Vb_0 = \epsilon + (r - Za_0)$ and $\|\eta\|^2 \le 2\|P_V\epsilon\|^2 + 2\|r - Za_0\|^2$. By the approximation property of splines, $\|r - Za_0\|^2 = O_p(np_1/K^{2d})$. Also, $E\|P_V\epsilon\|^2 = E(\epsilon^T P_V\epsilon) = \sigma^2\mathrm{tr}(P_V) = O(Kp_1 + p_2)$ and the lemma is proved by an application of Markov's inequality. □

We collect some miscellaneous results on bounding some terms used in the proof of Theorems 1 and 2 in the following lemma.

Lemma 2. *Following the notations used in Theorems 1 and 2, we have*

(i) $\|(\tilde{Y} - \tilde{X}\beta_0)^T\tilde{X}\| = O(\sqrt{np_2})$.
(ii) $\|\tilde{X}^T\tilde{X}/n - \Xi\| = o(1)$ *where $\|B\|$ for a matrix, B denotes its operator norm.*

Proof. (i) We first write down the decomposition

$$X = \Theta - G + G + U.$$

The above uppercase letters represent $n \times p_2$ matrices, and correspond to the decomposition in (4) evaluated at n observations. After projection, we have

$$\tilde{X} = \tilde{\Theta} - \tilde{G} + \tilde{G} + \tilde{U}.$$

Together with the decomposition

$$\tilde{Y} - \tilde{X}\beta_0 = \tilde{\epsilon} + \widetilde{(r - Za_0)}$$

(as in the proof of Theorem 2, $r = (r_1, \ldots, r_n)^T$ with $r_i = \sum_{j=1}^{p_1} W_{ij}\alpha_{0j}(t)$, a_0 contains the spline coefficients that achieve optimal approximation of $\alpha_{0j}(t)$, $1 \le j \le p_1$), the bound for $\|(\tilde{Y} - \tilde{X}\beta_0)^T\tilde{X}\|$ is obtained from the following estimates,

$$\|\epsilon^T Q_Z X\| = O(\sqrt{np_2}), \tag{7}$$

$$\|(r - Za_0)^T Q_Z(\Theta - G)\| = \sqrt{\frac{np_1^2 p_2}{K^{2d-1}}}, \tag{8}$$

$$\|(r - Za_0)^T Q_Z U\| = \sqrt{\frac{np_1 p_2}{K^{2d}}}, \tag{9}$$

$$\|(r - Za_0)^T Q_Z G\| = \sqrt{\frac{np_1}{K^{2d}}}\sqrt{\frac{np_1 p_2}{K^{2d_g}}}. \tag{10}$$

Proof of (7):

$$\begin{aligned} E\|\epsilon^T Q_Z X\|^2 &= E\epsilon^T Q_Z XX^T Q_Z\epsilon \\ &= O(\mathrm{tr}(Q_Z XX^T Q_Z)) \\ &= O(\mathrm{tr}(XX^T)) = O(np_2). \end{aligned}$$

Proof of (8):
We have $\|(r - Za_0)^T Q_Z(\Theta - G)\| \le \|(r - Za_0)^T(\Theta - G)\| + \|(r - Za_0)^T P_Z(\Theta - G)\|$. Since entries of $\Theta - G$ have mean zero and are orthogonal to $\mathcal{G}$ while entries of $r - Za_0$ are inside $\mathcal{G}$, we can obtain the bound

$$\|(r - Za_0)^T(\Theta - G)\|^2 = O\left(\frac{np_1 p_2}{K^{2d}}\right) \tag{11}$$

by considering its variance. For any fixed $1 \le j \le p_2$, let Θ_j be the j-th column of Θ and G_j the j-th column of G. We have

$$\begin{aligned}\|P_Z(\Theta_j - G_j)\|^2 &\le \|Z^T(\Theta_j - G_j)\|^2 \|(Z^TZ)^{-1}\| \\ &= O(\mathrm{tr}(ZZ^T))\|(Z^TZ)^{-1}\| \\ &= O(np_1)\cdot O(K/n) = O(p_1K).\end{aligned}$$

Since $\|r - Za_0\| = O(\sqrt{np_1/K^{2d}})$, the above implies $\|(r - Za_0)^T P_Z(\Theta - G)\|^2 = O(np_1/K^{2d})\cdot O(p_1p_2K) = O(np_1^2p_2/K^{2d-1})$.

Proof of (9):

Since $\|(r - Za_0)Q_Z\| \le \|r - Za_0\| = O(\sqrt{np_1/K^{2d}})$, we have $E\|(r - Za_0)Q_ZU\|^2 = O(np_1p_2/K^{2d})$, similar to (11).

Finally, (10) is obtained from $\|Q_ZG_j\| = O(\sqrt{np_1/K^{2dg}})$ by Condition (c7).

All the terms (7)–(10) are of order $O(\sqrt{np_2})$ by Condition (c8).

(ii) Using the decomposition $\tilde{X} = \Gamma - P_Z\Gamma + \tilde{G} + U - P_ZU$ where $\Gamma = \Theta - G$, we have that, using the results on eigenvalue convergence for the sample covariance matrix in [1],

$$\left\| \frac{(\Gamma + U)^T(\Gamma + U)}{n} - \Xi \right\| = O\left(\sqrt{\frac{p_2}{n}}\right) = o(1). \tag{12}$$

Denoting by Γ_j the j-th column of Γ, we also have the following bounds.

$$\begin{aligned}\frac{\Gamma_j^T P_Z \Gamma_j}{n} &= \frac{\Gamma_j^T Z(Z^TZ)^{-1}Z^T\Gamma_j}{n} \\ &\le \|\Gamma_j^T Z\|^2 \lambda_{\max}((Z^TZ)^{-1}/n) \\ &= O(\mathrm{tr}(ZZ^T))\lambda_{\max}((Z^TZ)^{-1}/n) \\ &= O(np_1)\cdot O(K/n^2) = O(Kp_1/n),\end{aligned}$$

which implies

$$\left\| \frac{\Gamma^T P_Z \Gamma}{n} \right\| = O\left(\frac{p_1p_2K}{n}\right) = o(1),$$

using the well-known inequality $\|W\| \le \mathrm{tr}(W)$ for any square matrix W.

By similar arguments, we have

$$\left\| \frac{U^T P_Z U}{n} \right\| = O\left(\frac{p_2}{n}\mathrm{tr}(P_Z)\right) = O\left(\frac{p_1p_2K}{n}\right) = o(1), \tag{13}$$

and finally

$$\left\| \frac{G^T Q_Z G}{n} \right\| = O\left(\frac{p_1p_2}{K^{2dg}}\right) = o(1) \tag{14}$$

by Condition (c7).

Other terms in $\|\tilde{X}^T\tilde{X}/n - \Xi\|$ can be bounded by the Cauchy–Schwartz inequality utilizing (12)–(14), resulting in some additional $o(1)$ terms, and Part (ii) of the lemma is proved. □

References

[1] Z. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, 2nd ed., in: Springer Series in Statistics, Springer, New York, London, 2009.
[2] C. De Boor, A Practical Guide to Splines, rev. ed., Springer-Verlag, New York, 2001.
[3] J.Q. Fan, R.Z. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (456) (2001) 1348–1360.
[4] J.Q. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, Annals of Statistics 32 (3) (2004) 928–961.
[5] P.J. Green, B.W. Silverman, Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, 1st ed., in: Monographs on Statistics and Applied Probability, Chapman & Hall, London, New York, 1994.
[6] T. Hastie, R. Tibshirani, Generalized Additive Models, 1st ed., in: Monographs on Statistics and Applied Probability, Chapman and Hall, London, New York, 1990.
[7] T. Hastie, R. Tibshirani, Varying-coefficient models, Journal of the Royal Statistical Society. Series B. Statistical Methodology 55 (4) (1993) 757–796.
[8] J.Z. Huang, Local asymptotics for polynomial spline regression, Annals of Statistics 31 (5) (2003) 1600–1635.
[9] J. Huang, J.L. Horowitz, S.G. Ma, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, Annals of Statistics 36 (2) (2008) 587–613.
[10] J. Huang, J.L. Horowitz, F. Wei, Variable selection in nonparametric additive models, Annals of Statistics 38 (4) (2010) 2282–2313.
[11] J.H.Z. Huang, C.O. Wu, L. Zhou, Polynomial spline estimation and inference for varying coefficient models with longitudinal data, Statistica Sinica 14 (3) (2004) 763–788.
[12] P.J. Huber, Robust regression: asymptotics, conjectures and Monte Carlo, Annals of Statistics 1 (5) (1973) 799–821.
[13] R. Li, H. Liang, Variable selection in semiparametric regression modeling, Annals of Statistics 36 (1) (2008) 261–286.

[14] G. Li, H. Peng, L.X. Zhu, Nonconcave penalized M-estimation with diverging number of parameters, Statistica Sinica 21 (2011) 391–420.
[15] L. Meier, S. Van de Geer, P. Buhlmann, High-dimensional additive modeling, Annals of Statistics 37 (6B) (2009) 3779–3821.
[16] S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency, Annals of Statistics 12 (4) (1984) 1298–1309.
[17] S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation, Annals of Statistics 13 (4) (1985) 1403–1417.
[18] P. Ravikumar, H. Liu, J. Lafferty, L. Wasserman, SpAM: sparse additive models, in: J. Platt, D. Koller, Y. Singer, S. Roweis (Eds.), Advances in Neural Information Processing Systems, vol. 20, MIT Press, Cambridge, MA, 2008, pp. 1201–1208.
[19] C. Stone, Additive regression and other nonparametric models, Annals of Statistics 13 (2) (1985) 689–705.
[20] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society. Series B. Statistical Methodology 58 (1) (1996) 267–288.
[21] L.F. Wang, H.Z. Li, J.H.Z. Huang, Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements, Journal of the American Statistical Association 103 (484) (2008) 1556–1569.
[22] H.S. Wang, B. Li, C.L. Leng, Shrinkage tuning parameter selection with a diverging number of parameters, Journal of the Royal Statistical Society. Series B. Statistical Methodology 71 (2009) 671–683.
[23] A.H. Welsh, On M-processes and M-estimation, Annals of Statistics 17 (1) (1989) 337–361.
[24] H.L. Xie, J. Huang, SCAD-penalized regression in high-dimensional partially linear models, Annals of Statistics 37 (2) (2009) 673–696.
[25] V.J. Yohai, R.A. Maronna, Asymptotic behavior of M-estimators for the linear model, Annals of Statistics 7 (2) (1979) 258–268.
[26] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (1) (2007) 19–35.
[27] C.H. Zhang, J. Huang, The sparsity and bias of the Lasso selection in high-dimensional linear regression, Annals of Statistics 36 (4) (2008) 1567–1594.
[28] H. Zou, The adaptive Lasso and its oracle properties, Journal of the American Statistical Association 101 (476) (2006) 1418–1429.
[29] H. Zou, R.Z. Li, One-step sparse estimates in nonconcave penalized likelihood models, Annals of Statistics 36 (4) (2008) 1509–1533.